DOCUMENT RESUME

ED 266 683 FL 015 539

AUTHOR Klein-Braley, Christine
TITLE Tests of Reduced Redundancy--Theory.
PUB DATE 85
NOTE 17p.; In; Language Testing in School. AFinLA Yearbook 1985 (see FL 015 537). See also FL 015 540.
PUB TYPE Reports - Research/Technical (143) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cloze Procedure; *Communicative Competence (Languages); *Dictation; Elementary Secondary Education; Foreign Countries; Information Theory; Language Proficiency; *Language Tests; Linguistic Theory; *Noise (Sound); *Redundancy; Second Language Learning; *Test Construction; Test Validity

ABSTRACT

One solution to the problems of discrete-point testing are tests of reduced redundancy, that is, tests that introduce interference into the communication through actual introduction of noise or through deletion of text and that ask the learner to perform in spite of the noise. Three test types were originally developed to operationalize the technique of reduced redundancy: the noise test, in which varying levels of acoustic noise were introduced during dictation of short sentences; partial dictation based on extremely difficult tests, omitting segments of two to ten words; and cloze tests, with systematic deletions of every fifth to tenth word. While each test type is found to have problems of validity or viability, the original theory remains convincing. A fourth test type, called the C-Test, has been developed as a better operationalization of the theory. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

ED266683

TESTS OF REDUCED REDUNDANCY - THEORY

Christine Klein-Braley (University of Duisburg)

0 Introduction

Scientific language testing is generally thought to have started in 1961 when Robert Lado published his book Language Testing, and John B Carroll his paper Fundamental considerations in testing for English language proficiency of foreign students. The tests which were developed as a result of these impulses consisted of multi-item subtests for the various aspects of language: grammar, vocabulary, phonology, reading and listening comprehension and so on. They conformed to all the demands of psychometrics. Since the items used were generally cast in multiple-choice form they were objective. Test development procedures using item analysis ensured high reliability. Correlations with other criteria such as teacher judgements showed their empirical validity.

Nevertheless in the early seventies linguists were becoming increasingly dissatisfied with this type of test. The multiple-choice item, they said, tested recognition rather than production, and there was no guarantee that a student who could identify the correct solution would necessarily be able to use the form in productive speech. Isolation of problems in individual items provided students with a clue as to what was being tested at any given moment. There seemed to be little theoretical justification for the implicit weightings accorded to the various subskills tested: a test which assigns twice as many points for vocabulary as for grammar makes the assumption that vocabulary is exactly twice as important in language, but there was no way of testing whether this assumption was true. In addition, the intercorrelations between subtests often turned out to be so high that - at least statistically - there was no real difference between the areas they were allegedly testing. The major criticism, however, was that such tests bore very little relationship to real-life language use: as Spolsky pointed out in 1973 they fail to take into account two vital truths about language, the fact that language is redundant, and the fact that it is creative.

1 Information theory and language

One of the solutions proposed to this problem was the construction of tests of reduced redundancy. The theoretical justification for such tests was put forward by Spolsky in 1973, and was based on Shannon and Weaver's communication theory (1949), nowadays generally referred to as information theory.

ED015539

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Tomala

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

This theory was originally developed under the auspices of the Bell Telephone Company, and is concerned with the efficiency with which a message is transmitted from a source to a receiver. The content of the message is the information, although it is important to realise that the concept of information, as it is used here, has nothing to do with the meaningfulness of the message. Information has to be looked on as something more abstract, like weight for instance. The information is encoded into symbols which may be numbers, letters, or, as in morse code, bleeps. At first it would seem that the ideal code for any message is one in which each symbol transfers the maximum amount of information. Numbers are like this. If a telephone number is to be transmitted, then every single digit is equally important: none can be missed out and none can be changed. No digit in the string can be reconstructed from knowledge of the ones preceding or following it. So numbers transmit the maximum amount of information. But because of this great care is needed, otherwise one can easily find that one has dialled Buenos Aires (from Germany 00541 ...) when one actually intended to dial London (00441 ...). Similarly, the letters B, P and V sound very alike in English and can easily be confused by the listener.

Any disturbance in communication is technically known as noise. Noise can occur at any point in the communication process: the typist can hit the wrong key, there can be atmospheric or electrical disturbances in the telephone line, the receiver can misread or mishear what arrives at his end. Noise is unpredictable and random and there is no way of telling whether a given message will be affected by it. But a message which arrives at the receiver so damaged that no sensible interpretation is possible is a waste of time and resources, and so noise must be countered if this is feasible.

For this reason communications engineers have to build safeguards into the system, and since it is very expensive to remove all technical sources of noise, they prefer to use more symbols than are strictly necessary and to ensure that the coding is performed in such a way that errors caused by noise can be minimised. For instance, both bank account numbers and the international standard book numbers include include check digits. When the numbers are entered into the computer some sort of arithmetical manipulation is performed, and if the result of the computation is not confirmed by the check digit, then the input is rejected as incorrect. If one is dictating a telegram, the post office asks one to use disambiguating names: "B for Bertha, P for Peter, V for Victory" etc. In this way, although it takes longer to dictate the telegram, one can be sure that the message has been taken down correctly.

The additions inserted into the message as safeguards are known as redundancy, and in terms of information theory redundancy is desirable. Shannon and Weaver's motivation for developing communication theory was to work out the mathematics of the process and thus to enable communications engineers to find the optimal balance between noise and redundancy in the transmission of messages - as much redundancy as necessary but as little as possible.

BEST COPY AVAILABLE

If we look at natural language in terms of information theory we find that it is a highly redundant system. Shannon estimated that printed English (in terms of letters) is about 75 % redundant. This means that the same amount of information could be conveyed by using only 25 % as many letters and spaces as at present. This is the technique adopted by the Speedwriting form of shorthand: most of the vowels are omitted, and some of the consonants, and yet the message can still be read:

gt a gd job & mo pay

say the advertisements for Speedwriting courses in the London tube.

As Spolsky points out, when one considers all the interferences that occur when natural language is used for communication, it is clear that only a redundant system would work. When we are confronted with bad handwriting, imperfect photocopies, noisy telephone lines, station loud speakers, it is redundancy which enables us to make sense of what we hear or read, even though some elements of the message are missing.

Moreover redundancy is present on a variety of different levels. Language is a very highly structured system consisting of interrelated subsystems, and we work with all these simultaneously. Just how much goes on when we are processing language becomes obvious if we look at the instructions which have to be given to that electronic idiot, the computer, if it is to understand and act on a simple English sentence (Diagram 1, taken from Winograd 1984). But we can perform this analysis without even thinking about it - because we know the rules of the language.

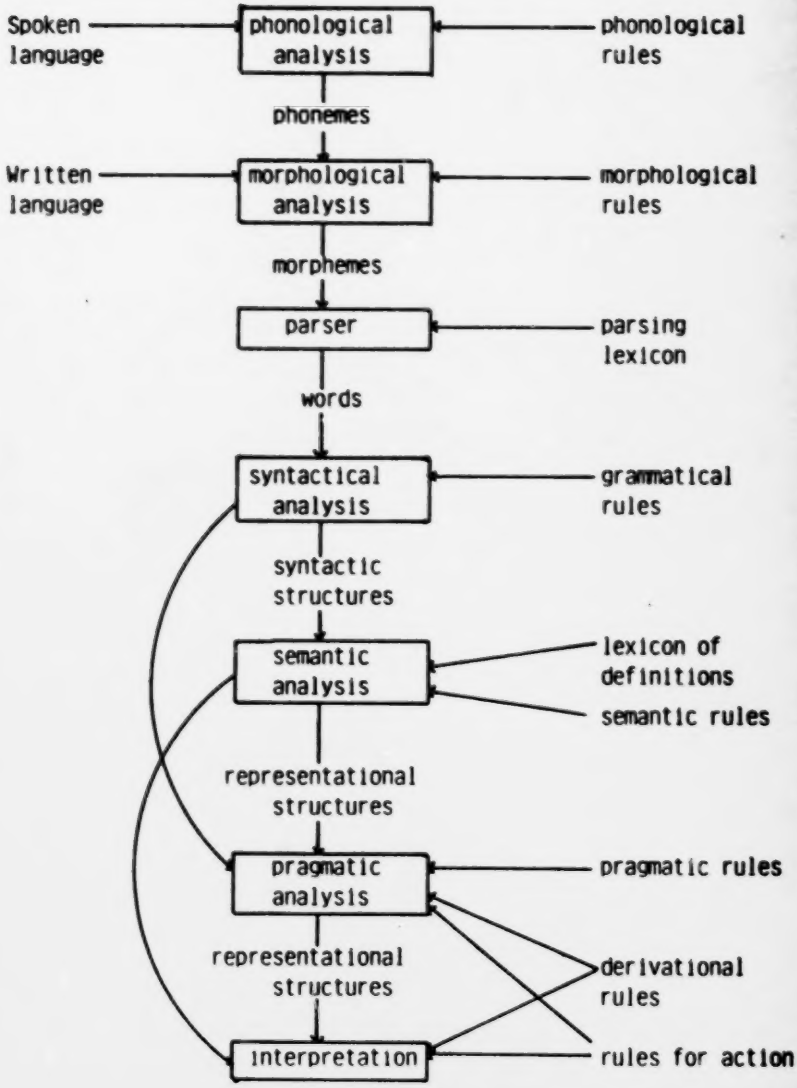
2 Redundancy and language testing

While native speakers of a language can use its redundancy to supply missing elements and thus restore the original message, non-natives need the full normal redundancy, and sometimes even more. Spolsky points out that when we talk to a foreigner we speak more slowly, more clearly, with added gesture and frequent repetition. He had been able to show (Spolsky et al. 1968) that in dictation tasks a clear distinction between natives and non-natives became evident as soon as noise was added. He explains that this is a result of:

> the non-native's inability to function with reduced redundancy, evidence that he cannot supply from his knowledge of the language the experience on which to base his guesses as to what is missing. In other words, the key thing missing is the richness of knowledge of probabilities - on all levels, phonological, grammatical, lexical, and semantic - in the language. (1973:170).

BEST COPY AVAILABLE

Diagram 1: Computer analysis of real language input



5 taken from Winograd 1984

In his view two implications for language testing follow from this:

1. It is problematical to try and decide that someone knows a language on the basis of his performance on discrete-point items since it is impossible to demonstrate that any one item is essential for successful communication.

2. Knowing a language certainly involves the ability to understand a distorted message, to make valid guesses about a certain percentage of omitted elements.

In order to make valid guesses about missing elements a language user has to know the rules of the language. In Chomky's terms he has to activate his competence (Chomsky 1965) in order to produce an acceptable performance. Spolsky therefore come to the conclusion that tests of reduced redundancy are techniques for making a learner "perform his competence".

Spolsky also points out that linguistic performance may be either active or passive: both the speaker and the listener are performing. Linguistic competence underlies both kinds of performance. The implications of this for language testing are important because it means that a test of the receptive skills can reveal competence just as well as a test of the productive skills.

A test of reduced redundancy, therefore, is one which introduces noise into the channel and asks the learner to perform in spite of this noise. The noise can be acoustic, but it can equally well be induced by deleting portions of a text.

3 Operationalisation of the theory

In order to convert a theory into a test a number of practical decisions must be taken, in other words the theory must be operationalized. A number of different operationalizations have been proposed in order to make use of reduced redundancy in language testing. They differ in a variety of aspects (see Table 1):

1. the definition of "text";
2. the definition of "noise";
3. the choice of channel;
4. the type of behaviour expected from the learner.

BEST COPY AVAILABLE

Table 1

Operationalization of reduced redundancy

TEST	text	noise	channel	examinee behaviour
Noise Test	single sentences	white noise at varying levels of intensity	acoustic	writing down what was heard
Partial dictation	2 longer texts	deletion of phrases selected by test constructor	acoustic/ visual	writing down what was heard in blanks (examinee provided with mutilated text)
Cloze test	1 longer text	deletion of words (every nth)	visual	filling in blanks in written text

However, in all cases they share the assumption that the performance shown on the text can be linked to the learner's competence, if not absolutely then at least in terms of rank ordering of the examinees, i.e. the examinee with the higher score has a more highly developed competence than the examinee with a lower score.

In the discussion which follows I should like to examine briefly three different operationalizations of the theory of reduced redundancy and discuss the implications of the results obtained with these tests for the original theory.

3.1. The Noise Test

The first studies with the Noise Test were reported in 1968 by Spolsky et al. The basic idea was to dictate short sentences to the examinees under varying conditions of noise. The examinees were to write down what they heard, or thought they heard.

A variety of preliminary experiments were conducted, and the final 1968 version of the test consisted of 50 sentences, controlled for syntactical complexity in terms of the number of transformations required, and controlled for vocabulary in terms of word frequency. White noise - which sounds like a constant hissing - was added acoustically at signal to noise ratios of 3 decibels (the highest noise level), 4, 7, 10 and 50 decibels. To avoid learning effects the highest noise levels were administered first. 5 points were given

for a correct response, 4 for a response with one phonemic error, 1 for a response with at least one word correct and 0 otherwise.

In all the early trials the test had proved to discriminate very well between native and non-native speakers (with the non-natives achieving lower scores). The final version was administered to 61 foreign students who entered Indiana University in 1986. The results of the Noise Test were correlated with the results of other English tests administered by the EFL department and a comparison of the scores with the levels of proficiency assigned by the EFL department (based on the test scores of their own tests and a short interview) showed that the Noise Test successfully identified the high and low scorers. There were some problems however in the middle range.

Table 2

Correlations between Noise Test and EFL tests

	IU Aural Comprehension	Objective	Essay
Noise test (84)	.66	.66	.51
IU Aural Comp		.61	.62
Essay			.65

In 1975 Gradman and Spolsky reported that a multiple-choice version of the Noise Test had been produced and that the correlation between the original and this version was .89 for one group of non-natives and .86 for a second. The test had also been administered to 26 Saudi-Arabian students who took the TOEFL at the same time. Here the correlation was .75 with the total TOEFL, and with the exception of English Structure and Writing (.44 and .33 respectively); all subtest correlations were above .80. In addition a correlation of .69 emerged between the Noise Test and the Ilyin Oral Interview. Similar correlations emerged for a second group of 25 Saudi students. For a group of 71 foreign students the correlation between the Noise Test and the Indiana University Placement Test was .56.

However there was a fly in the ointment. In the process of construct validation, i.e. in the attempt to relate test results and theory, the scores on the last 10 sentences, which were accompanied by a very high signal/noise ratio (50 db) and which therefore seemed to have no noise overlying them, were correlated with the other 40 test items. The correlation proved to be .93, "which may lead one to believe that as an overall measure of language proficiency, the noise test might just as well be given as a dictation test without the added distorting noise" (Gradman/Spolsky 1975: 65). This suspicion hardened when other noise levels were examined: only at the 7 db level and below did the distorting noise prove more advantageous to the native speaker.

BEST COPY AVAILABLE

Examination of the test sentences revealed that those sentences on which subjects did least well were the ones for which comprehension seemed to be most dependent on context. The researchers therefore concluded that in these cases reduced redundancy was indeed operating, but it was redundancy of context which was missing and which most hampered non-native performance. A contextualized version of the Noise Test was therefore developed in two parallel forms with 20 items each. The test material consisted of short dialogues and mini-lectures of the type which might be encountered in the classroom or in everyday life. This time pink noise was added.

The tests still discriminated well between natives and non-natives, and the natives had mean scores of 99 %. However, when the data were analysed in detail it was impossible to show that the noise had had any effect on the scores. The mean difference between scores with and without noise was only 3.9 points for Form A and a mere 1.6 points for Form B. The rank correlations for the 10 sentences with noise and the 10 without were .97 on Form A and .91 on Form B for the non-natives. Reluctantly, but honestly, the researchers came to the conclusion that: "the use of background noise seems to have had little effect on the measurement of overall proficiency for our subjects" (Gaies et al. 1977: 56).

We can therefore conclude that this attempt to operationalize reduced redundancy is not a viable one in its present form (1).

3.2 Partial dictation

Johansson's operationalisation of the concept of reduced redundancy was the partial dictation test. As he points out himself (1973b:2) it is closely related to two other measures thought to be tests of general language proficiency: dictation and cloze tests (2). Johansson reviews the work of Oller (1971) and Valette (1967), both of whom come to the conclusion that dictation is a satisfactory test of overall language proficiency, but he seems to feel that normal dictations are too easy for advanced students. Partial dictations, which are based on extremely difficult texts, are therefore a suitable technique.

Such tests have a number of advantages, claims Johansson. It is possible to concentrate on selected portions of the text. The passage is played from a tape recorder, which means that a variety of different voices and speech situations are possible. Students listen to the recordings only once, so that partial dictation is more economical.

Johansson describes two different techniques for developing his dictations. In the first case authentic material is recorded and then transcribed, in the second a suitable text is selected and then read on to tape by a native speaker. In both cases the test constructor decides what is to be omitted and ensures that the necessary pauses are patched into the tape immediately after the item deleted in the student text. The decision about what to delete is left to the test constructor's imagination. Technically it is advisable to delete segments of language which come before a natural pause in the text. Johansson's deletions vary between 2 and 10 words. Administration in

a language laboratory is advisable.

The tests are scored by assigning one point for every correct word restored. Spelling errors which do not interfere with understanding are not counted.

Johansson reports a correlation of .88 (N=80) between his two texts (parallel test reliability) and high correlations between these tests scores and other tests (Table 3):

Table 3

Correlations between partial dictations and other tests

English vocabulary	English vocabulary and grammar	English pronunciation	Swedish vocabulary	English speaking test
.80 (N=20)	.85 (N=20)	.78 (N=20)	.85 (N=20)	.82 (N=35)

However there are two puzzling things about these statistics. The first is why the full correlation matrix is not presented - this is normal practice since it enables the reader to examine patternings in the data for him/herself. Mean scores and standard deviations are also missing. The second curiosity is the inclusion of only 20, allegedly randomly selected, students in the correlations when 80 students took the test.

So far as the statistics go, this approach seems to work satisfactorily for these two texts and this sample. However there are problems. Spolsky et al. were working with a specific test, and at no point in their discussions do they imply that their results are to be generalized to other test versions. Johansson makes precisely this claim - "due to the high correlations with other measures of foreign language proficiency, it might be used as a global estimate of a subject's proficiency in a foreign language" (1973 b:42). This is a dangerous recommendation to make on the basis of only one test run and an analysis of only one quarter of the test subjects, particularly since there is no attempt made at any point to provide some sort of theoretical framework for the deletion techniques used.

Nor is any attempt made to link up the results obtained to a general theory of reduced redundancy. One possible approach, for instance, would have been to use the two other tests explicitly mentioned in the paper, cloze and dictation, as validation criteria. Or something along the lines of the multitrait-multimethod matrix would have been possible, using the integrative measures on the one hand and the discrete-point tests already available on the other.

In other words, there is no reason to assume at this point that partial dictation is a valid test of reduced redundancy, nor that

BEST COPY AVAILABLE

it measures general language proficiency. We just don't know!

3.3 Cloze tests

The most important and well-known operationalization of reduced redundancy is, of course, the cloze test. The test - and the name - were first used by Taylor (1953). It consists, as everyone probably knows, of a piece of prose in which every nth word been deleted (where n is a number between 5 and 10). The examinee restores the text by filling in the missing words. Scoring is either exact, so that only the words which the original author used are counted as correct, or it is acceptable, and in this technique any word which fits into the text semantically and syntactically scores a point.

The difference between the cloze test and other gapfilling procedures is the explit link-up between deletions and information theory. The systematic deletions are intendend to function in the same way as random noise in the communications system - Taylor's pilot studies had allegedly shown that systematic deletion was a good approximation to random deletion using random number tables.

Taylor's original aim was to develop a new readability technique, and in his experiments it was the texts, not the subjects, which were being tested. However he did, in fact, find significant differences between individuals and therefore suggested that cloze tests could be used to measure reading proficiency in the native language (L1).

In 1959 Carrol, Carton and Wilds experimented with cloze procedure for foreign language testing, but their results were negative. In 1968 Darnell obtained positive results with cloze tests for non-native speakers of English, but since his scoring procedures involved first testing a group of natives and then assigning points for the non-native responses on the basis of the frequency with which the particular response had been offered by the natives (clozentropy) this was not a technique which could be advocated for general use. However in 1971 Oller and Conrad performed an investigation using exact scoring, and produced results which showed that natives scored higher than non-natives and that the mean scores of the five non-native groups faithfully reflected their teaching classifications (from beginners to graduates in TESL). Furthermore the cloze scores showed a multiple correlation with the UCLA ELSPE test of .88. Oller and Conrad concluded: "This pilot study lends further support to the applicability of the cloze technique to the problems of ESL proficiency testing" (1971:192).

Oller continued to work on cloze tests. In 1972 he examined questions of difficulty levels and scoring techniques, and concluded that for all three levels of difficulty acceptable scoring was better: the correlations with the validation criteria were higher and the tests were more reliable. In 1973 he turned his attention to the question of what cloze tests measure:

> On the basis of incomplete information, the subject is required to project a word to fill in a blank and thereby complete a sentence. In doing so, both the so called "productive" (or "active")

> and "receptive" (or "passive") skills of a language must be utilized. The information provided in the cloze test allows the student by analysis to synthesize a greater whole. At the same time, the synthesis or projection may become part of the next analysis required to produce a subsequent synthesis.
>
> It is interesting to note that the process ... involves more than "passive" reading. By sampling the information that is present the subject formulates hypotheses, or expectations, about information that is to follow. By sampling subsequent sequences, he either confirms or disconfirms these expectations. If the expectations are disconfirmed they must be revised and new hypotheses must be formed ... I would like to suggest that they are ... manifestations of the ... underlying competence. (1973:114).

He coined the term pragmatic expectancy grammar, and linked it to Miller's concept of "chunking" information to explain the subjects' behaviour (Miller 1970:20 ff.). Miller had pointed out that more material could be retained in short-term memory if it could be "chunked" or structured, and had also shown that grammatical structuring was a very efficient method of "chunking". All this, of course, fits in very well with Spolsky's original theory for tests of reduced redundancy.

Empirically, cloze tests repeatedly showed excellent performances. As measured by KR 20 they were highly reliable, and they showed high correlations with the validation criteria (usually discrete-point tests). Furthermore the highest correlations were generally found between cloze tests and other holistic approaches to proficiency testing: listening comprehension tests and dictations, for instance. (For overview, see Oller 1979). It came to be generally accepted that cloze tests were reliable and valid tests of general language proficiency. They were recommended for use by teachers: they were easy to construct, to score, particularly by the exact method, and their reliability and validity were known to be high.

It is importance to realise, however, that the claims for high reliability and validity are based on theoretical assumptions:

1. The learner has reached a certain level of competence in the foreign language, i.e. has acquired a structured international grammar which is activated in all use of the language.

2. Cloze tests force the learner to use his internal grammar to restore missing elements.

3. The score obtained on a cloze test is a numerical estimate of the efficiency or quality of the learner's grammar.

This means that the text in a cloze test has a specific function: it is a sample of the language. The blanks are then a sample of the text, and these in their turn are intended to sample the learner's grammar. Only by making these theoretical assumptions can we legitimately state, as Oller, for instance, does, that the specific text involved is not important; any text will do (1979:364). This is also the reason that

BEST COPY AVAILABLE

deletion rate is not important: what is actually intended is random deletion, and random sampling is the best way of ensuring a representative sample of whatever we wish to measure.

The most obvious way of testing these assumptions is to examine the equivalence of cloze tests. If the theory is correct then the following statements should be true:

1. Random deletion of words and nth deletion of words produce equivalent samples of the text.

2. Tests constructed from the same texts but using different deletion rates or starting points should be equivalent.

3. The results obtained for the same sample of examinees using two different cloze tests should be equivalent.

Fairly recently, and unfortunately only after cloze tests had already achieved the important position in language testing that they hold today, two major investigations into these questions were reported: Alderson in 1978 and Klein-Braley in 1981. Their results showed that many of the claims put forward in favour of cloze tests were exaggerated.

To investigate question 1 Klein-Braley coded 13 texts into 12 word classes and then used a computer to construct all 45 cloze tests for each text which could be produced using the deletion rates 5, 6, 7, 8, 9 and 10 and all possible starting points. Statistical tests calculated on the basis of the two super-ordinate categories, structure and content words, showed that nth word deletion was not equivalent to random deletion. Indeed, empirical investigations showed that the actual difficulty of tests constructed from the same text varied according to the proportion of structure words deleted.

Both researchers examined the second question. Klein-Braley defined equivalence in terms of mean item difficulty and reliability coefficients. The results are shown in Table 4. For the same text difficulty varies between .36 and .65 (exact scoring) and reliability from .46 to .73 (exact scoring). For acceptable scoring the variability of difficulty is smaller (.73 to .87) but the reliability coefficients range from .38 to .62.

BEST COPY AVAILABLE

Table 4

Same text, different deletion rates

	Deletion	Mean item difficulty	Reliability
exact	8	.36	.46
	9.1	.56	.72
	9.3	.65	.73
	9.4	.61	.71
	10	.59	.62
accept.	8	.73	.80
	9.1	.79	.61
	9.3	.87	.38
	9.4	.83	.62
	10	.79	.60

Alderson defined equivalence in terms of validity coefficients: the correlations between the various cloze tests and the ELBA test. Again considerable differences were revealed. His easy text showed a correlation with ELBA of .59 at deletion rate 6 but of .70 for deletion rate 8; his medium text correlated .57 for deletion rate 6 but .86 for deletion rate 10.

In order to investigate question 3 Klein-Braley administered 6 different pairs of cloze tests to sample groups of students at the University of Duisburg (Table 5).

Table 5

Correlations between pairs of cloze tests
Administered to the same samples

	Exact scoring	Acceptable scoring
1 & 2	.55	.51
3 & 4	.39	.36
5 & 6	.48	.59
7 & 8	.58	.52
9 & 10	.70	.74
11 & 12	.64	.36

For acceptable scoring the correlations range from .36 to .75; for exact scoring from .39 to .70 (cf. Klein-Braley 1983:224).

These results, which represent only a very small part of the empirical findings of both researchers, offer no justification for the

notion that cloze tests are equivalent to each other. But the theory states that if cloze tests are operationalizations of reduced redundancy then they must be equivalent.

4 Conclusion

At the end of this review of tests of reduced redundancy where do we stand? We have a theory which states that one way to assess an individual's competence in a language, his internalized rule system, is to present him with a text where the redundancy has been reduced. In this way he is forced to activate all he knows about the language in order to restore the missing portions of the text. The better his competence, the more easily he will be able to do this. We can therefore draw conclusions about his competence from his performance on such tests.

On the other hand we have seen that out of three attempts to operationalise this theory in the form of tests, not one has proved satisfactory.

As Cronbach and Meehl (1955) point out:

The investigator whose prediction and data are discordant must make strategic decisions. His result can be interpreted in three ways:

1. The test does not measure the construct variable.
2. The theoretical network which generated the hypothesis is incorrect.
3. The experimental design failed to test the hypothesis correctly.

On other words, one can reject the theory together with the test. Or one can retain the theory and try to find a new and better operationalization. In Duisburg we still found the original theory convincing, and we therefore chose to use the third strategy. In 1981 Raatz and Klein-Braley first presented the C-Test to an IUS Symposium, and Raatz discusses our test of reduced redundancy in his paper (see Raatz, this volume).

Notes

1 It had been criticized as early as 1973 by Johansson, who particularly emphasised that the scoring technique was, in his opinion, unsatisfactory. His own noise test did not seem to function very well, but it is difficult to assess his results against those of Spolsky and his team since Johansson changed many of the determinants of the situation: he used lower signal/noise ratios (which theoretically ought to increase the difficulty) while at the same time constructing sentences deliberately very high in semantic and/or syntactic redundancy (which presumably reduce the difficulty). He also changed the scoring system. (cf. Johansson, 1973a).

BEST COPY AVAILABLE

2 It is difficult to see how dictation in the traditional form can be regarded as a test of reduced redundancy, although it is, of course, one of the major techniques advocated as a test of general proficiency.

Bibliography

Alderson, J.C. 1978. A study of the cloze procedure with native and non-native speakers of English. Unpublished PhD Dissertation, University of Edinburgh.

- 1983. "The cloze procedure and proficiency." In Oller, J.W.Jr. (ed) 1983. Issues in language testing research. Rowley, Mass: Newbury House. 205-217.

Carroll, J.B. 1961. "Fundamental considerations in testing for English language proficiency of foreign students." In Testing the English proficiency of foreign students. Washington D.C.: Center for Applied Linguistics. 30-40.

- , Carton, A.S. and Wilds, C.P. 1959. An investigation of cloze items in the measurement of achievement in foreign languages. College Entrance Examination Board Research and Development Report. Cambridge, Mass.: Laboratory for Research in Instruction, Harvard University.

Chomsky, N. 1965. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press.

Cronbach, L.J. and Meehl, P.E. 1955. "Construct validity in psychological tests." Psychological Bulletin 52: 281-302.

Darnell, D.K.1968. The development of an English language proficiency test of foreign students using a clozentropy procedure. Boulder, Co.: Department of Speech and Drama, University of Colorado.

Gaies, S.L., Gradman, H.L. and Spolsky, B. 1977. "Toward the measurement of functional proficiency: contextualization of the noise test." TESOL Quarterly 11,1: 51-57.

Gradman, H.L. and Spolsky, B. 1975. "Reduced redundancy testing: a progress report." In Jones, R.L. and Spolsky, B. (eds) 1975. Testing language proficiency. Arlington, Va.: Center for Applied Linguistics. 59-70.

Johansson, S. 1973a. "An evaluation of the noise test: a method for testing overall second language proficiency by perception under masking noise." IRAL 11,2: 107-132.

- 1973b. Partial dictation as a test of foreign language proficiency. Swedish-English contrastive studies 3. Lund University, Dept. of English.

Klein-Braley, C. 1981. Empirical investigations of cloze tests. An examination of the validity of cloze tests as tests of general language proficiency in English for German university students. PhD Dissertation, University of Duisburg.

- 1983. "A cloze is a cloze is a question." In Oller, J.W.Jr. (ed) 1983. Issues in language testing research. Rowley, Mass.: Newbury House. 218-228.

Lado, R. 1961. Language testing. London: Longman.

Miller, G.A. ^{2}1970. The psychology of communication. Harmondsworth: Penguin.

Oller, J.W.Jr. 1971. "Dictation as a device for testing foreign language proficiency." ELT 25: 254-259.

- 1972. "Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language." Modern Language Journal 56,2: 151-158.

- 1979. Language tests at school. London: Longman.

- and Conrad, C. 1971. "The cloze technique and ESL proficiency." Language Learning 21: 183-195.

Shannon, C.E. and Weaver, W. 1949. The mathematical theory of communication. Urbana: University of Illinois Press.

Spolsky, B. 1973. "What does it mean to know a language; or how do you get someone to perform his competence?" In Oller, J.W.Jr. and Richards, J.C. (eds) 1973. Focus on the learner. Rowley, Mass.: Newbury House. 164-176.

- , Bengt, S.M., Sako, E.W. and Aterburn, C. 1968. "Preliminary studies in the development of techniques for testing overall second language proficiency." In Upshur, J.A. and Fata, J. (eds) 1968. Problems in foreign language testing. Language Learning Special Issue No. 3: 79-103.

Taylor, W.L. 1953. "Cloze procedure - a new tool for measuring readability." Journalism Quarterly 30: 414-438.

Valette, R. 1967. Modern language testing. New York: Harcourt Brace and World.

Winograd, T. 1984. "Software für Sprachverarbeitung." Spektrum der Wissenschaften. Nov. 1984: 88-132.

BEST COPY AVAILABLE